基于贝叶斯的 ROC 分析方法在心理测量研究中的应用

刘雨晴　　李慧玲　　周强*

（温州医科大学精神医学学院应用心理系　温州　325000）

摘要： ROC（receiver operating characteristic ）分析是诊断研究中一种重要且应用广泛的方法。虽然近年来其广泛应用于诊断研究，但国内尚未见该方法用于心理测量研究中。而 ROC 分析方法中，本文主要介绍基于贝叶斯理论的 ROC 分析（BROC）在心理测量中的应用。基于此，本文首先回顾介绍 ROC 分析方法的发展及演变，然后梳理 BROC 在心理测量中的应用，并进行实例模拟，最后展望其在心理测量领域的应用前景。

关键词：ROC 分析；贝叶斯；心理测量；诊断测验；准确性评估；

1. 引言

心理学研究常利用生理或心理指标来量化心理状态和/或特质，从而达到预测和控制相关行为的目的，因此指标的准确性评估是保证预测结果的重要前提。问卷测量及行为实验等是研究相关心理与行为的常用测量工具，如使用大五人格量表来反映人格特质（Lui et al.，2020），利用贝克抑郁量表来测量评估个体的抑郁情况（Macchi et al.，2020），利用 ERP 来研究个体内部心理状态等（Cui et al.，2021）。心理测量工具的准确性是研究有效的重要前提。但以往心理学研究中常用信效度来反映测量工具的有效性，其结果较为单一，无法直观反映结果的预测价值，且无法直接比较不同测量工具之间的准确性，因此如何用更好的方法来评估心理测量工具的准确性迫在眉睫。

ROC 分析最先在信号检测论（SDT）中被提出，最早用于雷达监测，后用于研究感觉阈限（如听觉、视觉和触觉）等行为反应；如今已被广泛应用于分析心理学和神经科学实验（Sumner et al.，2019），以及其他各种不同的领域，如医学诊断、机器学习等（Obuchowski &Bullen，2018 ；Ma et al.，2019），可由 R 中的 plotROC 包实现（Sachs，2017）。近年来，国外应用 ROC 分析方法对心理测量工具进行准确性评估的研究越来越多（Ruddy et al.，2018；Bowers et al.，2019；Thapa et al.，2020），主要是将时间依赖相关的 ROC 分析

*通讯作者邮箱：zq@wmu. edu. cn
*本文系国家社会科学基金项目“20BSH047”的研究成果之一

（tROC）和基于贝叶斯的 ROC 分析（BROC） 在诊断研究中的使用方法用于心理学研究中。如 Levis 和 Sun （2020）利用 ROC 分析方法来比较抑郁症筛查量表 PHQ－2、PHQ－9 及联合诊断之间的评估准确性。ROC 分析方法还一般通过二分类转换，寻找最佳临界点，从而获得更多所需要的信息。如，Richardson（2018）等在研究智能手机使用时，利用 ROC 分析来获得智能手机使用量表（PSUS）的最佳阈值，通过计算 AUC 来评估 PSUS 的准确性，并利用 cut－off 点去寻找连续性结果的最佳临界值。

随着计算机技术的发展，在医学诊断研究中，ROC 分析方法不仅实现了对金标准条件的放宽，从二分金标准、等级金标准到无金标准，而且还能在研究过程中将更多协变量的影响考虑在内，如时间依赖相关的 ROC 分析、基于贝叶斯原理的无金标准 ROC 分析方法等。后者能够在无金标准下进行诊断评估，彻底摆脱过去 ROC 分析必须基于金标准存在的壁垒，从而为缺乏金标准的一些研究提供可能性，这为 ROC 在心理测量准确性评估中的应用提供了启示。与此同时，虽然 ROC 分析方法早已涉足心理学领域，但它在国内心理测量中的应用尚未得到广泛应用。

基于此，本文先简单归纳现存的 ROC 分析方法，尤其是基于贝叶斯的 ROC 分析（BROC），然后总结它在心理测量领域的具体应用，并就其在心理学领域的进一步发展提出展望。旨在将 BROC 分析方法 “移植”到心理测量领域，从而拓宽其在心理学领域尤其是心理测量领域的应用范围。

2. ROC方法的常见方法介绍

ROC曲线是一个以1－特异性（specificity）为横坐标，敏感性（sensitivity）为纵坐标的曲线关系图(见图1)，主要利用曲线的临界值（cut－off point）和曲线下面积（AUC，area under curves）来反应诊断结果（Mandrekar et al.， 2010）。

曲线的临界值（cut－off point）,即曲线拐点处的正切值，在临床研究中常选择最大约登指数（Youden index）所对应的临界值，即最佳cut-off值，作为将测试结果划分为阳性和阴性的依据。约登指数表示诊断方法准确区分患者与非患者的总能力（灵敏度与特异度之和减去1），指数越大说明筛查实验的效果越好，真实性越大（Martínez–Camblor et al.，2019）。曲线下面积（AUC）的形式定义是：$\mathrm{AUC} = \int_0^1 y(x)dx$，即对所有可能的特异性值进行检验的敏感度平均值，AUC越高的测试被认为是准确性越好。但AUC的指标变化敏感性低，因此单靠AUC的比较无法直接得出结论，故还需要结合参考敏感性和特异性值(Janssens & Martens，2020) 。

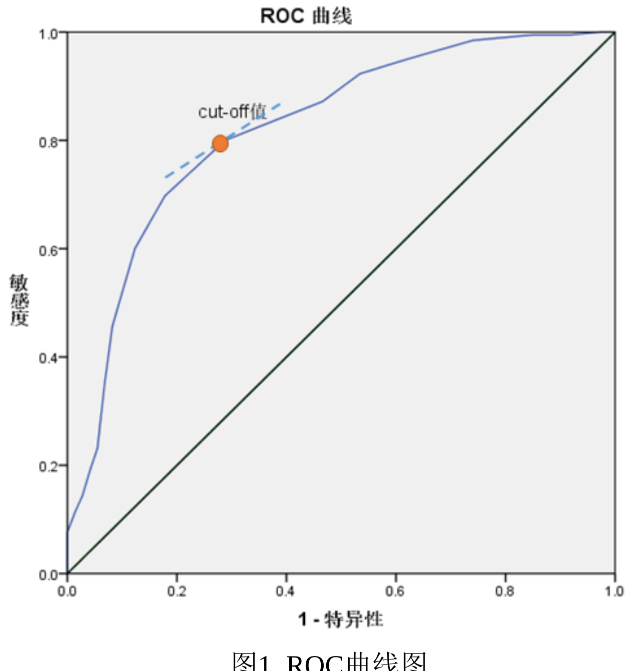


图1. ROC曲线图

传统 ROC 分析应用于诊断评估时通常使用 Yerushalmy 模式，核心是将所测结果与金标准做比较。因此前提是存在一个可靠、稳定的二分金标准，否则将无法计算其灵敏度与特异度，从而无法评价和判断准确性。尽管金标准对 ROC 分析而言至关重要，但要获得一个稳定、合适的二分金标准并不容易。临床上很多疾病的金标准并非二分变量，而是等级或连续变量；此外有些金标准获取成本极高，过程繁琐复杂，或不符合伦理道德要求，甚至暂时没有成熟的金标准，由此可见对金标准的严格要求极大地限制 ROC 分析方法的应用（王肖南，2019）。为解决此类问题，有研究者利用专家经验将等级变量主观转化为二分变量。如 Numan 等利用专家经验将三类合并为两类（Numan et al.， 2019），以方便进行 ROC 分析，但主观转换造成的误差较大；再如陈卫中等将等级变量分为多组二分变量，分别两两比较，从而为等级变量的金标准研究提供方向，但这本质上仅是扩展曲线下面积的应用，并无法利用 ROC 曲线的其它信息，如 cut–off 值等（陈卫中，2012）。事实上早在上世纪末就有研究者（Peng et al.，1996）将贝叶斯理论引入 ROC 分析方法中，实现无金标准的 ROC 分析。与过去的 Yerushalmy 模式不同，该方法主要利用贝叶斯理论，不局限于寻找金标准，而强调收集先验信息，再结合临床经验获得的对疾病有效的相关信息，与此同时还能将多个协变量的影响考虑在内，从而对其后验分布进行有效的估计。Flor 等研究表明贝叶斯估计方法优于传统的频率估计方法（Flor et al.，2020）。

根据金标准的特征不同，本文总结出以下三种常用新方法：等级金标准条件下的ROC

曲线分析方法，时间依赖相关的ROC曲线分析方法，无金标准条件下的ROC曲线分析方法。以下是对这三种类型方法的详细介绍，包括金标准主要特征、临床应用及评价。

2. 1 基于等级金标准条件下的ROC曲线分析方法

等级金标准条件下的ROC分析方法不仅可用来对等级或连续数据的诊断方法进行准确度评价，还可根据要求将等级变量转化为二分变量。其基本过程是先将各等级状态下的数据两两比较，并分别计算曲线下的面积（AUC），最后比较AUC，以达到评价的效果（陈卫中，2012；Obuchowski et al.，2005）。例如陈卫中在评价氧化低密度脂蛋白ELISA 检测试剂盒在冠心病诊断中的诊断价值中，按金标准将被试分为三类状态（有病，无病，可疑）。AUC估计与互相比较可通过R软件中的nonbinROC包（Paul Nguyen et al.，2007）实现，更多R包详情及操作方法详见该研究。

2. 2 时间依赖相关（time–dependent）的ROC曲线分析方法（在本文用‘tROC’代替）

tROC分析方法主要是通过拓展敏感性与特异性的概念，并观察它在每个时间点的疾病状态，从而产生不同的敏感性和特异性，以此获得一个与时间相关的ROC曲线图（Kamarudin et al.，2017）。此外还可直接得到不同时间点的AUC，从而获得关于AUC(t)的函数图，以便直观有效地比较同一测量指标以及不同测量指标之间在不同观测时间的准确性。此方法最早由Heagerty和Zheng（2005）提出，其研究发现可利用每个时间点t的累计敏感性与动态特异性（C/D）、事件敏感性与动态特异性(I/D)以及事件敏感性与静态特异性(I/S)等三种不同定义评估上述时间观测事件的敏感性与特异性，从而适用于不同的情境。

tROC分析中可观察个体疾病的连续状态，增加个体发病时间的信息，还能在时间点间构建ROC曲线，并比较各测量指标的预测能力。这在临床上有颇为广泛的应用，例如Suzuki等通过生存分析评估SIS和mGPS对预后的影响，利用随时间变化的受试者操作特征分析，比较感兴趣的各个评分对预后的影响（Suzuki et al.，2018）。再如Lima 等利用ROC法结合振荡梯度自旋回波(OGSE)和脉冲梯度自旋回波(PGSE)的不同扩散时间，探讨ADC值变化在头颈部肿瘤良恶性鉴别中的应用（Lima et al.，2019 ）。tROC可通过R包实现，具体可参考（Díaz–Coto et al.，2020）的研究。

2.3 基于贝叶斯理论（Bayesian theory）的无金标准ROC分析方法（下文用‘BROC’代替）

上述ROC分析方法依赖于金标准，但在临床实践中许多疾病的金标准获取成本颇高，甚至缺乏金标准。基于此，Peng （1996）等提出将贝叶斯理论引入ROC分析中，即在无金标准条件下仍可考虑多个协变量的影响，且可计算不同协变量影响下的ROC曲线下面积

（AUC），从而比较诊断准确性。

贝叶斯理论与频率统计不同，认为概率是主观的，并主张将个体经验信息作为重要部分来推导后验分布。基本原理是先根据模型的样本似然函数，结合参数的先验分布，从而推导出后验分布，即由先验概率乘以似然值而获得后验概率。近年来随着计算机技术的进步，贝叶斯理论被广泛应用于许多领域，尤其是在医学的诊断研究和心理测量工具准确性评估中（Arora ＆ Thorlund, 2019；Goyal ＆ Yolcu, 2019；Park ＆ Lee, 2019 ）。在诊断准确性评估研究中，首先需要根据目标人群相关信息，确定先验信息，这是第一步也是最为关键的一步；其次通过似然函数对参数的先验分布进行调整，从而推导出后验分布，实现对相关诊断方法灵敏度和特异度的估计，更多详情可参见相关资料（McClean et al.，2014）。

因此对无金标准诊断实验评价而言，只要有一定的实验诊断先验信息，再结合一些并非金标准但临床证实有效的现时观测数据，就可以通过贝叶斯理论推导出诊断实验评价指标的后验分布，从而摆脱对金标准的依赖。例如Amini（2020）等利用贝叶斯潜在分类模型（LCMs）以联系诊断测试观察结果与潜伏疾病状态，在无完全准确疾病状态分类的情况下评估诊断准确性。除此之外，BROC尚能同时考虑多个协变量的影响。相比前面几种方法，其在本质上摆脱ROC分析方法受金标准的限制，从而拓展了ROC分析方法在医学、心理学、计算机等多个领域的应用。如Zi－Hui Tang（2014）的研究，利用贝叶斯模型评估压力反射敏感性（BRS）进而预测心血管自主神经病变（CAN）。在CAN无金标准的前提下，选取2092疑似病例，将年龄、血压等作为协变量，以BRS为诊断标准，使用贝叶斯潜在类模型来评估BRS的敏感性和特异性。结果发现BRS在CAN诊断试验中具有较高的敏感性和特异性，具有一定的参考价值，提示BRS检验是诊断CAN的有效工具（Zi－Hui Tang et al.，2014）。其实早在2012年QiuWang等人就提出将BROC分析应用在教育学与心理学当中，结合贝叶斯层次模型和接受者操作特征分析（BROC）来评估兴趣强度（IS）和兴趣分化（ID）如何预测低社会经济地位（SES）青年的兴趣－专业一致性（IMC））（QiuWang et al.，2012）。当然，本文仅介绍贝叶斯方法在无金标准条件下的诊断应用，事实上贝叶斯理论的应用远不止此，还包括深度学习、潜变量建模、多水平结构建模、实验数据分析等。随着交叉学科思想的进一步深入，将贝叶斯理论与相关研究领域结合的应用也越来越多。当然贝叶斯理论模型也并非完美，因为它过分强调经验的重要性，容易造成主观偏差，从而影响结果准确性。

总的来说，ROC 分析是一种全面的，且准确评估诊断准确性和预测价值的方法，广泛应用于医学和心理学。近年来，结合实际需求，在传统二分金标准条件下 ROC 分析的基础上，发展出适用于不同临床条件下的适用方法，研究结果也充分证明其合理性。

3. BROC分析方法在心理测量中的应用

如引言部分所述，ROC 分析方法虽然在心理学领域已经应用颇多年，但是其仍然限于研究感知觉阈限及认知加工等领域，从根本上来说其局限于二分金标准。但随着计算机科学的进一步发展，ROC 分析方法已有较大的新进展，尤其是在诊断研究方面。因而本文实质上关注“移植”到心理学领域中的 ROC 分析方法，尤其是基于贝叶斯原理的 ROC 分析方法给心理学研究带来的启示，并就已有的相关研究进行总结梳理。

3. 1 量化某种心理测量工具的预测价值（准确性）

在心理的临床应用中，常需要根据测量结果对数据进行分类，从而有助于做出是或否、有或无的判断。例如心理学的相关选拔测试中时，需要对连续性结果数据进行分类，从而做出是否符合企业要求的判断；而其在心理疾病的测量中也是尤为重要，如根据抑郁量表的得分多少，最终将其与特定值比较，从而做出是否患抑郁症的判断。在过去的研究中，我们常使用平均数或者中位数进行二分转换，而在心理疾病诊断中，例如抑郁量表得分中，我们常将其与固定的得分作比较，对其做出分类。但事实上，这样分类的准确性并无法对其进行评估。而 BROC 分析可根据曲线上拐点的正切值得到获得阈限值（cut－off），并参考约登指数找到最佳 cut－off 值从而将连续变量的结果划分为两类。Cut-off 值作为诊断研究中多年来最佳分类指标，将其应用于心理学的二分类转换中是具有十分大潜力的。例如抑郁症，焦虑症，强迫症等评估中可在测试中得出 ROC 曲线，根据 cut－off 值，结合医生的意见即可做出是否有抑郁症的诊断。除临床诊断外，ROC 分析还适用于心理普测。例如 Battaglia 等利用 ROC 分析方法，获得 ESAS physical、psychological 和 global 子量表的最佳分界点并比较 KTR 与 ICD－10 诊断和 DCPR 诊断的 ESAS 评分（Battaglia et al., 2020)；再如 Thapa 等利用 ROC 曲线分析的方法判断自杀意念和自杀企图中的三维心理痛苦（DPPS）作为检测高自杀风险抑郁症患者的有效筛查量表的准确性（Thapa et al.，2020）。

问卷法作为心理学研究中最常用的测量工具之一，其广泛应用于心理特质测量以及心理疾病的诊断研究中。而其本身的准确性以及预测价值的评估是保证测量有效的重要前提。例如利用大五人格问卷来预测人格特征，测量个体情感障碍的人格特征易感性（Wilks et al.，2020），以及预测主观幸福感和心理幸福感（Anglim et al.，2020）。过去的研究多采用信效度检验，通过信效度系数来反应其有效性和适用性，如使用赫龙巴赫系数（a）反应其信度，但此方法无法直观反应其准确性与预测价值。而 BROC 分析方法可以通过曲线下面积（AUC）直接量化其在该研究中的准确性，弥补了传统信效度检验方法的不足。例如 Zeinab 等在判

断人格特质对心理问题的行为预测研究中，通过 BROC 分析确定人格特征对伊朗成年人常见心理问题的预测价值，利用 BROC 分析方法，分别获得三种问卷的 ROC 曲线，并比较曲线下的面积，得出神经质对于预测常见心理问题有良好的价值（Zeinab et al.，2017）。再如 Kassing 等使用 BROC 分析方法来利用儿童早期的行为问题去预测成人的信念（Kassing et al.，2019）。Lin GM（2020）等则利用 BROC 分析来判断相关机器学习模型对军事人员自杀意念预测的准确性的好坏评价。

此外，BROC 分析方法不仅仅适用于问卷研究中，其同样适用于实验研究中，如磁共振与脑研究等等。Stevens 等利用 BROC 研究功能性磁共振成像(fMRI)在脑肿瘤术前定位中的可靠性(Stevens et al., 2016)，再如 Raes 等利用 BROC 评估经颅磁刺激（TMS）的准确性并利用贝叶斯潜在类别模型诊断马脊髓功能障碍(Raes et al., 2020)，Gu 等也利用 BROC 分析判断 MRI 对疾病的诊断性能(Gu，2019)。在诸多心理疾病的诊断研究中，过去常将 ICD 作为诊断标准，但诸多心理特质的测量结果是不具备金标准的，而 BROC 分析可以实现在无金标准的条件下对其进行准确性评估，这一方法的应用为心理学测量工具的准确性评估打开了一扇大门。

3. 2 比较不同条件下的测量工具

ROC 分析可以通过获得曲线下面积（AUC），以便对不同的筛查或诊断试验进行有意义的比较（Walker，2019）。众所周知，对于同一心理特质我们常通过不同的测量工具来进行研究，例如测量心理渴求的相关问卷根据总结发现达到五份以上，如《依赖程度量表问卷》、《使用药物渴求问卷》、《毒品复吸高危量表》、《成瘾物质渴求与自动化行为反应量表》等。除此之外，不同外在条件下，同一工具测量的结果可能会出现差异。因此单单利用信效度一个指标来对其有效性下直接的结论是片面且缺乏科学性的，且利用唯一的固定值来对其结果进行二分类转换也会对结果造成偏差，而利用 ROC 分析可以很好的回避此问题。我们在此基础上梳理发现，不同条件下的测量工具的比较主要包括不同被试、不同时间、不同测量工具之间的不同。

在心理学研究中，比较不同样本之间对同一心理因素的差异对理论和实践具有重大的意义。过去对于不同被试样本同一心理特征的差异比较通常是利用参数检验来实现，但它要求数据呈正态分布。而 BROC 分析对数据分布形态无要求，可直接利用曲线图比较不同样本之间的差异，并通过 ROC 曲线图更加直观清晰的呈现结果。对于某些心理特质，不同人群之间可能就会存在不同的差异，那么其在不同被试人群中的准确性就可能存在差异。例如不同职业从事人员事业倦怠可能不同即可通过 ROC 分析来比较。

ROC 曲线方法可以独立比较两个或多个测量工具的准确性。同一个心理现象或者心理因素由于理论基础和维度不同所使用的测量工具可能存在差异。不同的研究者对同一心理问题的研究可能采用不同的量表，但是很少有人将不同的量表之间进行准确性的比较，因此用于同一心理测量的问卷之间本身可能就可能存在差异性，从而导致形成不同的研究结果，不利于后人重复研究结果和进行元分析。所以研究者存在比较不同测量工具准确性的需求，通过 BROC（贝叶斯的 ROC）分析可实现此目标，在无金标准的情况下，独立比较不同测量工具之间的差异，并评价其准确性。如 Chenneville 等人利用 ROC 分析探讨比较 PHQ 和 CES－D 对艾滋病毒感染者青少年抑郁症筛查的效用(Chenneville et al., 2019)；再如 Hartung 等人利用 ROC 分析方法来评估医院焦虑抑郁量表(HADS)和 9 项患者健康问卷(PHQ－9)作为筛查癌症患者抑郁的工具的有效性比较(Hartung et al., 2017)。

上述介绍的皆是 BROC 分析方法在心理学中的应用，实际上 tROC 分析也是心理学纵向研究中重要可取的方法。其不仅仅可以单独比较某一测量工具的准确性，更是能够考虑时间等协变量因素的影响。tROC 目前常用于生存分析中，尤其是对癌症晚期病人生存时间的预测上。遗憾的是过去虽然关于其在生存分析中具有较多的研究，但也受限于此，极少有研究者将其用于其他纵向研究中。其在心理学中的应用更是少之又少，但其在心理学领域的潜力不可小估。如 Liu 等人采用 tROC 分析来评估肌电活动随时间变化的动态预测性能，并通过 ROC 曲线获得最佳 cut-off 值，将强直性和阶段性肌电活动分为轻度和重度两类（Liu et. al.，2019）。再如测量心理渴求的量表可能在戒毒人员于戒毒所的戒毒时长不同，其测量效果可能存在差异，这亦与我们接下来的研究紧密相关。

综上所述，ROC 分析方法是一种适用于心理学，医学等诸多领域的研究方法。近年来 ROC 分析方法在心理学中的应用不仅限于信息加工，还用于心理测量工具的比较与评价，但总体来说其应用在心理学领域方兴未艾。系统全面地梳理 ROC 分析方法在心理测量准确性评估领域的新进展有利于全面推动该方法的应用。

4. 实例演示

为更好的说明 ROC 分析方法在心理测量领域的应用，本文利用 OpenBUGS 软件，采用人工数据，模拟 BROC 分析方法在心理测量中的应用实操。BROC 分析首先需要选择合适的模型，再根据实际的需要选择并设置不同的参数，然后验证模型，最后利用软件获得其 ROC 曲线以及 AUC、cut-off 值等等。本次实验模拟的是对 100 名受试者的海洛因成瘾情况，获得判断是否成瘾最佳的阈限值，以及量化本次研究的准确性。先让 100 名被试完成《海洛因依赖量表》，并记录得分，以下是分析过程。

本次模拟假设有 100 个受试者：$i=1,2,\ldots,100$。其中受试者的成瘾问卷分数计为 Y_i，年龄等人口学变量计为X_i。假设第 i 个人的真实情况d_i（成瘾=1，不成瘾=0）且人在成瘾的情况下和不成瘾的情况下 测试得到的问卷分数是连续变量且其得分的分布都是正态的，并且是两个不同的正态分布，即：

$$Y|d=0\sim N(\alpha,\tau) \qquad Y|d=1\sim N(\alpha'=\alpha+\beta,\tau)$$

由上可知d_i 实际上是二项分布，即$d_i\sim Bern(\pi_i)$是$d_i=1$ 的概率（这个人是否成瘾），加入人口学等协变量的影响即：$logit(\pi_i)=\eta+\psi*X_i$。在贝叶斯模型下，我们给予这些参数适当的先验分布（prior）:$\alpha\sim N(0,1)$ $\beta\sim N(0,1)$ $\eta\sim N(0,1)$ $\psi\sim N(0,1)$ 正态分布$\tau\sim gamma(0.001,0.001)$ gamma 分布（因为 tau 是正数）。假设我们选择"eta"、"psi"等参数，使用 gibs 抽样的方法通过反复迭代来让参数收敛，此次模拟迭代三次，其结果如图 2 所示，其相互重叠，说明迭代效果良好。此外计算模型各参数的秩相关结果发现相关系数趋于 0，说明模型正常，结果如图 3 所示。最后获得 ROC 曲线图（见图 4）及相关信息。

当然，本次模拟是解决在无金标准的前提下，对连续变量进行 ROC 分析的过程。其不仅可以应用于问卷数据结果的准确性测量以及结果分类以及行为实验结果。此次模拟的具体代码可联系通讯作者。

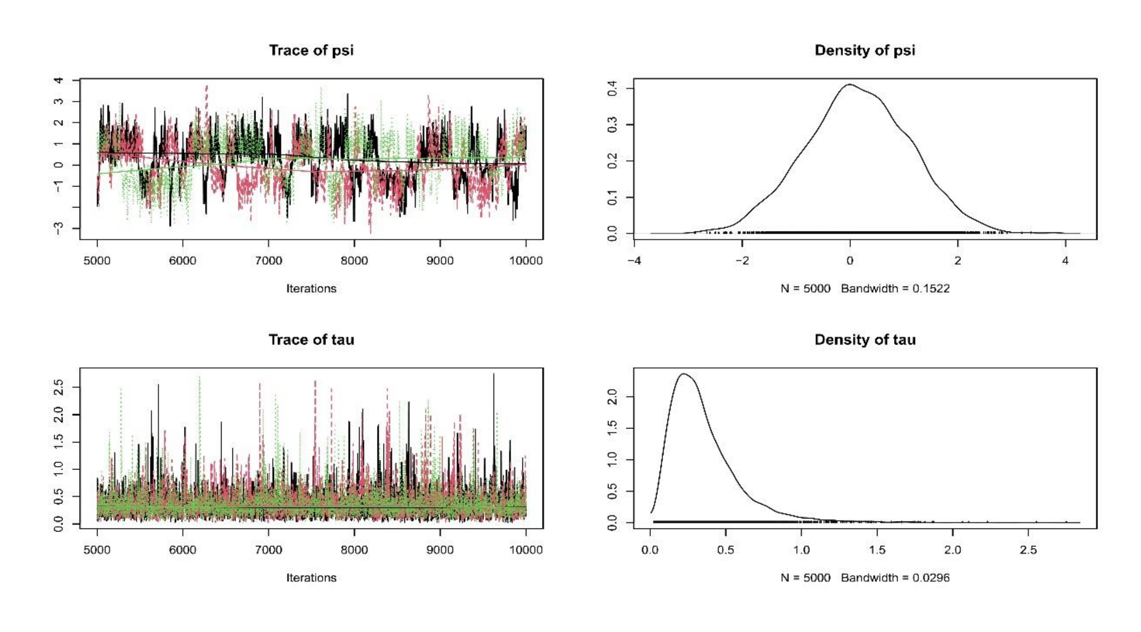


图 2 参数拟合度和概率密度函数图

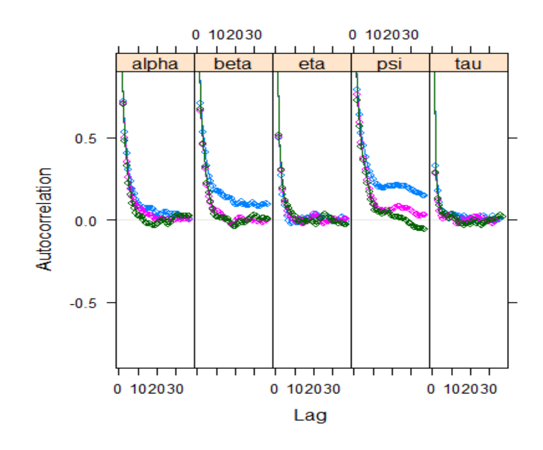


图 3　模型参数秩相关结果

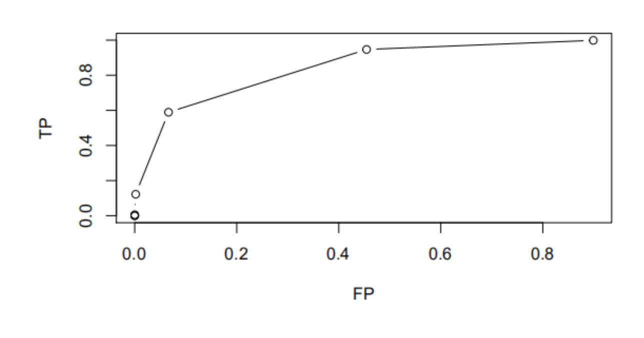


图 4　ROC 结果图

5. 总结与展望

自 ROC 分析方法被应用到诊断研究中以来，随着临床工作的需要和研究技术的进步，方法上屡有突破，但是国内尚缺乏对 ROC 具体方法进行的系统综述。而近年来虽偶有将 ROC 分析方法应用于心理学的研究，但并未见对其在心理研究中具体应用的梳理总结。为此本文不仅总结整理 ROC 分析方法的具体进展，还梳理它在心理学研究中的具体应用。

本文首先整理不同条件下的 ROC 分析方法的应用，并就其实现方式做了简单的介绍，然后就该方法在心理学中的应用做详细的阐述总结。如前文所述，整体而言 ROC 分析方法本身已经较为成熟，近年来它在心理测量工具准确性评估中的应用也越来越多。作者总结梳理具体方法上的进展以及在心理学中的具体应用，认为其应用于心理测量评估领域尚存在一定的问题需要解决。

首先， ROC 分析在心理测量工具评估中的使用价值需要更多的实际研究支撑。ROC 分析方法最大的优势是可以获得 ROC 曲线图，从而直观的独立比较其准确性差异。过去的 ROC 分析方法作为评估诊断价值的良好手段主要用于医学领域中，尽管近年来国外逐渐有研究出

现心理测量评估中的应用研究，但基于心理指标与生理指标的特点不同，ROC 在心理测量中的作用仍然需要更多的实证研究来证明。此外在使用 ROC 分析方法的同时还应该结合具体的实际情况，尤其是当用于心理疾病的诊断研究时，应该要综合考虑医生的主观判断，做出最后的判断。

其次，BROC 的应用价值值得进一步深入探讨。BROC 分析方法是各个 ROC 分析方法中限制条件最为宽松的，无需金标准即可评估测量工具准确性。而此方法的提出和使用为其在心理学中的应用打下良好的基础。例如目前在物质成瘾的研究中基于心理渴求测量方式主要为问卷，脑电等相对客观的研究工具仍需要问卷结果予以锚定。但由于测量心理渴求的问卷不同，结果可能会因为测量方式的不同而存在差异，而在不同戒毒时间段不同的测量工具的准确性亦可能存在差异。此外，缺乏一种可以量化心理渴求感程度，并且做出是否有“心瘾”判断的方法。我们的后续研究将会与此相关，进一步将 ROC 分析应用于心理渴求感的诊断研究。由此可见，利用 BROC 分析来评估心理渴求相关测量工具的准确性具有重要的理论和实践意义。此外，在心理学研究中主要通过问卷法和实验法来测量心理现象与行为活动，而 BROC 分析可以在无金标准的条件下独立计算比较量表和实验结果的有效性。

除此之外，ROC 分析可以融合机器学习、计算精神病学等交叉学科进行研究。近年来随着计算机科学的进一步发展，机器学习和计算精神病学逐渐成为研究热点，不仅广泛应用于图像识别、语言处理和数据挖掘，医疗领域等（Komura& Ishikawa，2019；Goecks & Jalili.，2020；Kan，2017；Crawley & Zhang.，2020），还在心理测量领域成为高级心理过程的研究工具（Bleidorn & Hopwood，2018；Shatte & Hutchinson，2019）。在机器学习的过程中评估模型准确性，并做出判断是必不可少的步骤，而这一步骤可以通过 ROC 分析来实现，其中 AUC 是机器学习中一种重要的性能评价准则，广泛应用于类别不平衡学习、代价敏感学习、排序学习等诸多学习任务（Dwyer & Falkai，2018）。总的来说 ROC 分析的应用范围仍然值得推广，本身具备有不可替代的作用。ROC 分析本身就像催化剂一样，能够应用于各个需要测量准确度的领域，并且由于它本身操作简单，结果却精确丰富，能够为诸多研究增添色彩。

参考文献：

陈卫中, 张菊英. (2012). 金标准为等级变量时诊断试验的评价及其在冠心病诊断试验中的应用. *中国卫生统计*, 29(2),172–174.

王肖南，周晓华，刘强，高颖. (2019). 无金标准下两种诊断方法准确度的贝叶斯估计. *中国卫生统计*, 36(05), 653 - 657.

宫学宇, 汪帅, 焦奥, 华向东.（2020）. 基于凝血酶原时间、纤维蛋白原和血小板平均容积的PFM评分系统对晚期胰腺癌病人生存的预测价值. *腹部外科, 33 (05)*，359–375.

Amini, M., Kazemnejad, A., Zayeri, F., Montazeri, A., Rasekhi, A., Amirian, A., & Kariman, N. (2019). Diagnostic accuracy of maternal serum multiple marker screening for early detection of gestational diabetes mellitus in the absence of a gold standard test. *BMC Pregnancy Childbirth, 20(1),* 375 –384.

Anglim, J., Horwood, S., Smillie, L. D, Marrero RJ, & Wood, J. K. (2020). Predicting psychological and subjective well–being from personality: A meta–analysis. *Psychological Bulletin,146(4),* 279 –323.

Arora, P., Thorlund, K., Brenner, D. R.,&Andrews, J. R. (2019). Comparative accuracy of typhoid diagnostic tools: A Bayesian latent–class network analysis. *PLOS Neglected Tropical Diseases. 13(5)*，1–23.

Bowers A.J., Zhou X. (2019). Receiver Operating Characteristic (ROC) Area Under the Curve (AUC): A Diagnostic Measure for Evaluating the Accuracy of Predictors of Education Outcomes. *Journal of Education for Students Placed at Risk. 24(1),* 20–46.

Battaglia Y., Zerbinati L., Piazza G., Martino E., Provenzano M., Esposito P., Massarenti S., Andreucci M., Storari A., Grassi L.(2020). Screening Performance of Edmonton Symptom Assessment System in Kidney Transplant Recipients. *Journal of Clinical Medicine. 9(4),* 995.

Blanche P, Dartigues J.F., Jacqmin–Gadda H. (2013). Review and comparison of ROC curve estimators for a time–dependent outcome with marker–dependent censoring. *Biometrical Journal. 55(5),* 687–704.

Bleidorn W, Hopwood C. J. (2019). Using Machine Learning to Advance Personality Assessment and Theory. *Personality and Social Psychology Review. 23(2),* 190–203.

Crawley D, Zhang L, Jones EJH, Ahmad J, Oakley B, San José Cáceres A, Charman T, Buitelaar JK, Murphy DGM, Chatham C, den Ouden H, Loth E; EU-AIMS LEAP group.(2020) Modeling flexible behavior in childhood to adulthood shows age-dependent learning mechanisms and less optimal learning in autism in each age group. *PLOS Biology. 18(10).*

Chenneville T, Gabbidon K, Drake H, Rodriguez C. (2019). Comparison of the utility of the PHQ and CES–D for depression screening among youth with HIV in an integrated care setting. *Journal of Affective Disorders,* 140–144.

Cui L, Dong X, Zhang S. (2021).ERP evidence for emotional sensitivity in social anxiety. *Journal of Affective Disorders. 279,* 361–367

Díaz Coto , Martínez Camblor P, Pérez Fernández S.(2020). Smooth ROC time: an R package for time–dependent ROC curve estimation. *Computational Statistics. 35(3),* 1231–1251.

Dwyer D.B., Falkai P., Koutsouleris N. (2018). Machine Learning Approaches for Clinical Psychology and Psychiatry. *Annual Review of Clinical Psychology. 14(1),* 91–118

Flor M., Weiß M., Selhorst T., Müller Graf C., Greiner M. (2020). Comparison of Bayesian and frequentist methods for prevalence estimation under misclassification. *BMC Public Health. 20(1),* 1135.

Goecks J, Jalili V, Heiser L.M, Gray J.W. (2020). How Machine Learning Will Transform Biomedicine. *Cell. 181(1),* 92–101.

Goyal A, Yolcu Y.U., Goyal A., Kerezoudis P., Brown D. A., Graffeo C.S., Goncalves S., Burns T.C., Parney I.F. (2019).The T2–FLAIR–mismatch sign as an imaging biomarker for IDH and 1p/19q status in diffuse low–grade gliomas: a systematic review with a Bayesian approach to evaluation of diagnostic test performance. *Neurosurg Focus. 47(6),* 13.

Hartung T.J., Friedrich M., Johansen C., Wittchen H.U., Faller H, Koch U., Brähler E., Härter M., Keller M., Schulz H., Wegscheider K., Weis J., Mehnert A. (2017). The Hospital Anxiety and Depression Scale (HADS) and the 9–item Patient Health Questionnaire (PHQ–9) as screening instruments for depression in patients with cancer. *Cancer. 123(21),* 4236–4243。

Heagerty , Zheng .(2005). Survival model predictive accuracy and ROC curves. *Biometrics. 61(1),* 92–105.

Janssens A.C.J.W., Martens F.K. (2020). Reflection on modern methods: Revisiting the area under the ROC Curve. *International Journal of Epidemiology. 49(4),* 1397–1403.

Martínez Camblor P., Pardo Fernández J.C. (2019). The Youden Index in the Generalized Receiver Operating Characteristic Curve Context. *International Journal of Biostatistics, 15(1).*

Kamarudin A.N., Cox T., Kolamunnage Dona R. (2017). Time–dependent ROC curve analysis in medical research: current methods and applications. *BMC Medical Research Methodology. 17(1),* 53.

Kassing F., Godwin J., Lochman J. E., Coie J.D. (2019). Conduct Problems Prevention Research Group. Using Early Childhood Behavior Problems to Predict Adult Convictions. *Journal of Abnormal Child Psychology. l47(5),* 765–778.

Komura D, Ishikawa S. (2019). Machine learning approaches for pathologic diagnosis. *Virchows Arch. 475(2),* 131–138.

Kan A. (2017). Machine learning applications in cell image analysis. *Immunology and Cell Biology. 95(6),* 525–530.

Lin G.M., Nagamine M., Yang S.N., Tai Y.M., Lin C., Sato H. (2020). Machine Learning Based Suicide Ideation Prediction for Military Personnel. *IEEE J Biomed Health Inform. 24(7),* 1907–1916.

Lui P.P., Samuel D.B., Rollock D., Leong F.T.L., Chang E.C. (2020). Measurement Invariance of the Five Factor Model of Personality: Facet–Level Analyses Among Euro and Asian Americans. *Assessment. 27(5),* 887–902.

Mandrekar J.N. (2010). Simple statistical measures for diagnostic accuracy assessment. *Journal of Thoracic Oncology.5(6),* 763–764.

Macchi C., Favero C., Ceresa A., Vigna L., Conti D.M., Pesatori A.C., Racagni G., Corsini A., Ferri N., Sirtori C.R., Buoli M., Bollati V., Ruscica M. (2020).Depression and cardiovascular risk–association among Beck Depression Inventory, PCSK9 levels and insulin resistance. *Cardiovasc Diabetol. 19(1),* 187.

McClean G., Riding N.R., Pieles G., Watt V., Adamuz C., Sharma S., George K.P., Oxborough D., Wilson M.G. (2019). Diagnostic accuracy and Bayesian analysis of new international ECG recommendations in paediatric athletes. *Heart. 105(2),* 152–159.

Ma Y, Ji .J., Huang Y., Gao H., Li Z., Dong W., Zhou S., Zhu Y., Dang W., Zhou T., Yu H., Yu B., Long Y., Liu L., Sachs G., Yu X. (2019). Implementing machine learning in bipolar diagnosis in China. *Translational Psychiatryy. 9(1),* 305.

Numan T., van den Boogaard M., Kamper A.M., Rood P.J.T., Peelen L.M, Slooter A.J.C. (2019).Dutch Delirium Detection Study Group. Delirium detection using relative delta power based on 1–minute single–channel EEG: a multicentre study. *British journal of anaesthesia. 122(1),* 60–68.

Obuchowski N.A. (2005). Estimating and Comparing Diagnostic Tests' Accuracy When the Gold Standard is not Binary. *Statistics in Medicine,20,* 3261–3278.

Obuchowski N.A., Bullen J.A. (2018) Receiver operating characteristic (ROC) curves: review of methods with applications in diagnostic medicine. *Physics in Medicine and Biology. 63(7).*

Peng F., Hall W. J. (1996). Analysis of ROC curves using Markov–chain Monte Carlo methods. *Medical decision making. 16(4),* 404–11.

Paul Nguyen. (2007). nonbinROC: Software for Evaluating Diagnostic Accuracies with Non–Binary Gold Standards. *Journal of Statistical Software,21(10)* , 1–10.

Park M.H., Lee S.H., Ko Y.H., Kim Y.K., Han K.M., Jeong H.G., Han C. (2019). Usefulness of the 15–item geriatric depression scale (GDS–15) for classifying minor and major depressive disorders among community–dwelling elders. *Journal of Affective Disorders. 259,* 370–375.

QiuWang, M. A. D. A. (2005) Applying Bayesian Modeling and Receiver Operating Characteristic Methodologies for Test Utility Analysis. *Educational and Psychological Measurement 73(2),* 275–292.

Ruddy, Jonah; Ciancio, Dennis; Skinner, Christopher H., Blonder, Megan (2018). Receiver Operating Characteristic Analysis of Oral Reading Fluency Predicting Broad Reading Scores. *Contemporary School Psychology*.

Richardson M., Hussain Z., Griffiths M.D. (2018) Problematic smartphone use, nature connectedness, and anxiety. *Journal of Behavioral Addictions. 7(1),* 109–116.

Raes E., Buczinski S., Dumoulin M., Deprez P., Van Ham L., van Loon G., Pardon B.. (2020).Accuracy of transcranial magnetic stimulation and a Bayesian latent class model for diagnosis of spinal cord dysfunction in horses. *Journal of Veterinary Internal Medicine. 34(2),* 964–971.

Suzuki Y., Okabayashi K., Hasegawa H., Tsuruta M., Shigeta K., Kondo T., Kitagawa Y. (2018). Comparison of Preoperative Inflammation–based Prognostic Scores in Patients With Colorectal Cancer. *Annals of Surgery. 267(3),* 527–531.

Stevens M. T., Clarke D. B., Stroink G., Beyea S. D., D'Arcy R. C. (2016). Improving fMRI reliability in presurgical mapping for brain tumours. *Journal of neurology, neurosurgery, and psychiatry. 87(3),* 267–74.

Shatte A. B. R., Hutchinson D.M., Teague S.J. (2019). Machine learning in mental health: a scoping review of methods and applications. *Psychology Medicine. 49(9),* 1426–1448

Sumner C. J., Sumner S. (2020).Signal detection: applying analysis methods from psychology to animal behaviour. *Philosophical transactions - Royal Society. Biological sciences. 375*.

Sachs M.C. (2017). plotROC: A Tool for Plotting ROC Curves. *Journal of Statistical Software*.

Thapa S., Sun H., Pokhrel G., Wang B., Dahal S., Yu S. (2020). Performance of Distress Thermometer and Associated Factors of Psychological Distress among Chinese Cancer Patients. *Journal Oncology*. 1-8.

Tang Z. H., Zeng F., Yu X., Zhou L. (2014). Bayesian estimation of cardiovascular autonomic neuropathy diagnostic test based on baroreflex sensitivity in the absence of a gold standard. *International Journal of Cardiology. 171(3),* 78–80.

Wilks Z, Perkins A.M., Cooper A., Pliszka B., Cleare A.J., Young A.H. (2020). Relationship of a big five personality questionnaire to the symptoms of affective disorders. *Journal Affect Disorder. 277,* 14–20.

Zeinab Alizadeh A, B. (2017). The predictive value of personality traits for psychological problems (stress, anxiety and depression): Results from a large population-based study. *Journal of Epidemiology and Global Health.8,* 124-133.

作者贡献声明[i]：

刘雨晴：负责提出问题，撰写论文

李慧玲、周强：论文最终版本修订

New Progress of ROC and Its Application in Psychometric Accuracy Assessment

Yuqing-Liu Hui-ling li Qiang zhou *

(Wenzhou medical university department of applied psychology, Wenzhou , 325000)

Abstract：ROC (Receiver Operating Characteristic) analysis is an important method widely applied in Diagnostic research. Although this method has been widely used in Diagnostic research in recent years, it has not been applied in psychological measurement in China. In the ROC analysis method, this paper mainly introduces the application of ROC analysis (BROC) based on Bayesian theory in psychometric measurement. Based on this, we not only review the concept of ROC analysis and its important indicators, but also summarize the Grade variable Gold Standard, Time Dependence Correlation and No Gold Standard. Consequently, the application value in psychological measurement require reviewed and prospected.

- Key words：ROC analysis；Bayes；Diagnostic test；psychological measurement；Accuracy assessment；
